Supplemental Methods and Analyses

Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species

Dorothea Lindtke[1,2], Zachariah Gompert[3], Christian Lexer[2], and C. Alex Buerkle[1]

[1] Department of Botany, University of Wyoming, Laramie, WY 82071, USA

[2] Unit of Ecology and Evolution, Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland

[3] Department of Biology and Ecology Center, Utah State University, Logan, UT 84322, USA

Corresponding author: Dorothea Lindtke
1000 E. University Ave.
Department of Botany, 3165
University of Wyoming
Laramie, WY 82071, USA
dlindtke@uwyo.edu
Fax: 307-766-2851

Keywords: paternity, parentage, admixture, reproductive isolation, Bayesian inference, next-generation sequencing

Running title: Hybridization frequency in *Populus*

Estimation of genetic ancestry

***Model description*:** Details on the model used to infer genetic ancestry of population reference samples, mothers and progeny, and the unsampled fathers are given below. Figure 2 (main article) shows a graphical representation of the full model specified in Equation 1 (main article). The model was written in `C`, using the GNU Scientific Library (Galassi *et al.*, 2009) and HDF5 (The HDF5 Group, 2010). `C` source code is available at Dryad (doi:10.5061/dryad.kh7sc).

We use a Bayesian method related to the approach implemented in the software `structure` (Pritchard *et al.*, 2000; Falush *et al.*, 2003), but with the difference that it works with genotype uncertainty arising from sequence data with limited coverage, estimates ancestry at both allele copies jointly, and can make use of family data. As described in the main article, the probability of observing the genotype $\mathbf{g}$ is conditional on the unknown population of origin $\mathbf{z}$ of the alleles that form the genotype, and the unknown allele frequencies $\mathbf{p}$ in the source populations, $P(\mathbf{g}|\mathbf{z},\mathbf{p})$. We use genotype likelihoods, $L(\mathbf{g}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{g})$ rather than raw sequence data $\mathbf{x}$ as model input. The genotype likelihoods were pre-calculated using `bcftools`, taking into account the number of reads, allelic state, and read specific error rate $\boldsymbol{\epsilon}$, given by or computed from the sequence data $\mathbf{x}$ (see main article; Li, 2011). The genotype likelihoods were normalized to sum to 1.

As we restrict our model to work with bi-allelic loci and diploid individuals, four different genotypic states $g_{ij} \in \{00, 01, 10, 11\}$ are allowed for each individual j and locus i (where 0 denotes the reference allele, and 1 the alternative allele). We can calculate the probability of the genotypic state as the product over the probabilities for allelic states of the first and the second allele copy, conditional on $\mathbf{z}$ and $\mathbf{p}$. This corresponds to a draw from a Bernoulli distribution (or a Binomial with a single draw) for each of the two allele copies with probability equal to the allele frequency of the alternative allele in the population of

origin k:

$$P(g_{ij}|\mathbf{z}_{ij},\mathbf{p}_i)=\prod_k\prod_a\begin{cases}p_{ik}^{g_{ija}}(1-p_{ik})^{1-g_{ija}} & \text{when } k=z_{ija},\\ 1 & \text{when } k\neq z_{ija}.\end{cases} \tag{2}$$

The population of origin of allele copy $a\in\{1,2\}$ is described by ancestry $z_{ija}\in\{1,\ldots,K\}$, and $g_{ija}\in\{0,1\}$ gives the state of the first or second allele copy in g_{ij}.

Although it is convenient to work with locus-specific ancestry z_{ija} for each allele copy separately in (2), addressing locus-specific ancestry conditional on genome-wide admixture can provide additional information by considering the diploid genotype. Therefore we calculated the probability for locus-specific ancestry jointly for both allele copies by working with ancestry genotypes $\mathbf{z}_{ij}$ rather than ancestry for each allele copy z_{ija} separately. The ancestry genotype $\mathbf{z}_{ij}$ can be seen as a $K\times K$ matrix with all its elements set to zero except the element at row $k=z_{ij1}$ and column $k'=z_{ij2}$ set to one. This matrix is used to describe all $K\times K$ possible ancestral genotypes for a given number of source populations K. The probability of the locus-specific ancestry genotype $\mathbf{z}_{ij}$ is then calculated conditional on the genome-wide admixture class matrix $\mathbf{Q}_j$ of individual j. $\mathbf{Q}_j$ is another $K\times K$ matrix that gives the prior probabilities for genome-wide admixture, or genome composition, for each of the possible states of $\mathbf{z}_{ij}$, with all elements in $\mathbf{Q}_j$ summing to 1 (see main article). If the gametic phase is unknown, the resulting matrix will be symmetrical above and below the main diagonal, with elements on or off the diagonal giving probabilities for intra-source or inter-source ancestry, respectively. The probability for locus-specific ancestry conditional on genome-wide admixture follows a categorical distribution (or a multinomial distribution with one draw) and is given by

$$P(z_{ij_{kk'}}=1|\mathbf{Q}_j)=Q_{j_{kk'}} \tag{3}$$

with k and k' giving the row and column of the $\mathbf{z}_{ij}$ and the $\mathbf{Q}_j$ matrix. The genome-wide admixture proportion $\mathbf{q}$ is not included as a model parameter but is calculated marginally

from $\mathbf{Q}$ within each iteration as

$$q_{j_k} = \frac{1}{2}\left(\sum_{s=1}^{K} Q_{j_{ks}} + \sum_{t=1}^{K} Q_{j_{tk}}\right). \tag{4}$$

We specify the prior probability for the genome-wide admixture matrix $\mathbf{Q}$ for mothers ($\mathbf{Q}_m$), fathers ($\mathbf{Q}_f$), and population reference samples ($\mathbf{Q}_r$) with a Dirichlet distribution with parameter vector $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{KK})$. To assign the same prior probability to each ancestral genotypic class, we specify identical values for all $\gamma_{kk'}$. This is appropriate when assuming that neither individuals with ancestry from a single population, nor hybrids dominate the hybrid zone. The size of $\gamma_{kk'}$ corresponds to the expected amount of admixture between the genotypic classes, where very small values indicate that most individuals have ancestry mainly from one ancestry class (i.e. pure species or F_1 hybrids in our analysis with two species). The hyperparameter $\gamma_{kk'}$ is drawn from a Uniform distribution $\gamma_{kk'} \sim \text{Uniform}(0, 10]$.

When working with family data, the prior on $\mathbf{Q}$ for the progeny ($\mathbf{Q}_p$) can be calculated directly from the admixture proportions of its parents with an expected value of $\boldsymbol{\nu} = (\nu_{11}, \ldots, \nu_{KK})$ and precision scalar β, with $\boldsymbol{\nu}$ being a function of the genome-wide admixture of the parents, $\boldsymbol{\nu} = f(\mathbf{Q}_m, \mathbf{Q}_f)$. The elements of $\boldsymbol{\nu}$ are calculated independently for each offspring $j = p$ from the admixture proportions of its mother, $\mathbf{q}_m^{(j)}$, and its father, $\mathbf{q}_f^{(j)}$, as

$$\nu_{j_{kk'}} = \frac{1}{2}\left(q_{m_k}^{(j)} q_{f_{k'}}^{(j)} + q_{m_{k'}}^{(j)} q_{f_k}^{(j)}\right) \tag{5}$$

for unphased data, ensuring matrix symmetry. We specify the precision of the prior on $\mathbf{Q}_p$ with $\beta \sim \text{Uniform}(0, 500)$ identical for all progeny. This leads to an informative Dirichlet prior for $\mathbf{Q}_p$ specified with $\boldsymbol{\nu}\beta = (\nu_{11}\beta, \ldots, \nu_{KK}\beta)$. A large β increases the prior information on ancestry for the offspring, given ancestry of its parents (Equation 21). The precision of the prior information also depends on the admixture proportion of the parents due to the multiplication $\boldsymbol{\nu}\beta$. For species parents, most weight of the prior will be on only one of

the elements in $\boldsymbol{\nu}$, whereas the weight will be more evenly distributed across elements for admixed parents. A large β (and thus precision of the prior) will therefore also reflect a large genetic map size, as the variance in the progeny's expected ancestry, given ancestry of its parents, will decrease with an increasing number of independent chromosome blocks (affected by the number of chromosomes, genetic map size of chromosomes, and recombination rate).

The probability of the unobserved allele frequency p_{ik} of locus i in source population k is calculated assuming an F-model, where the population allele frequency p_{ik} is the result of divergence from allele frequency π_i of a common ancestral population. We draw p_{ik} from a Beta distribution with shape parameters π_i and $(1-\pi_i)$, both multiplied by $(1/F_k - 1)$. F_k can be seen as a measure of genetic divergence of population k away from the ancestral population, analogous to the fixation index F_{ST}:

$$P(p_{ik}|\pi_i, F_k) \sim \mathrm{Beta}\left(\pi_i\left(\frac{1}{F_k}-1\right), (1-\pi_i)\left(\frac{1}{F_k}-1\right)\right). \tag{6}$$

The allele frequencies π_i are obtained from a symmetrical Beta distribution, $P(\pi_i|\alpha) \sim \mathrm{Beta}(\alpha, \alpha)$. The hyperparameter α can be seen as a measure of genetic diversity in the ancestral population, and is drawn from a Uniform distribution $\alpha \sim \mathrm{Uniform}(0, 10000]$. F_k is assigned an uninformative prior $F_k \sim \mathrm{Beta}(1,1)$.

***MCMC initialization*:** The MCMC process is started by assigning random values to parameters at the lowest level of hierarchy, and by subsequently initializing the other parameters accordingly. To accelerate convergence of the MCMC algorithm, we provide starting values for $\mathbf{q}$. Initial values for q_{jk} were obtained by concatenating genotype likelihoods (normalized to sum to 1) into a single point estimate (i.e. $L(g_{ij}=01)+L(g_{ij}=10)+2L(g_{ij}=11)$) per locus and individual, and running a PCA (`prcomp`-function in `R`; R Development Core Team, 2012) on those estimates. The first five principal components were then subjected to a linear discriminant analysis (`lda`-function in `R`, MASS package), where grouping was given by the assignments to K clusters obtained using the `kmeans`-function (`R`, MASS package).

For fathers, all q_{jk} were initialized as $1/K$.

***MCMC settings and updates*:** To achieve good mixing behavior of the chains, we ran an adaptation phase of 10,000 iterations with widened proposal intervals for $\boldsymbol{\gamma}$ and $\mathbf{Q}_f$. After adaptation, we continued with the actual MCMC algorithm comprising 100,000 iterations by using the proposal distributions specified below. We discarded the first 50,000 iterations as burn-in, and then took 2,000 samples with a thinning interval of 25. As sample distributions for $\mathbf{Q}_f$ were temporarily stuck in some cases, we combined samples from 10 independent chains for our final results. For simulated data, 3 chains were combined, and some of the proposal intervals were adapted to match the properties of the data. For all analyses, we considered only $K = 2$, corresponding to hybridization between two parental species, according to results from previous work (Lindtke *et al.*, 2012).

The model parameters were updated as follows:

1. Update $\mathbf{g}$ (sampled from the full distribution)
2. Update $\mathbf{z}$ (sampled from the full distribution)
3. Update $\mathbf{p}$ (Gibbs sampling)
4. Update $\boldsymbol{\pi}$ (Metropolis-Hastings)
5. Update $\boldsymbol{F}$ (Metropolis-Hastings)
6. Update α (Metropolis-Hastings)
7. Update $\mathbf{Q}$ (Gibbs sampling or Metropolis-Hastings), computed marginal $\mathbf{q}$
8. Update β (Metropolis-Hastings)
9. Update $\boldsymbol{\gamma}$ (Metropolis-Hastings)

Below, we specify the updates in more detail:

1. Update $\mathbf{g}$ (for all $j \neq f$):

$$P(g_{ij} = \{g_{ij1}, g_{ij2}\} | L(g_{ij}|x_{ij}), \mathbf{z}_{ij}, \mathbf{p}_i) =$$

$$\frac{L(g_{ij}=\{g_{ij1},g_{ij2}\}|x_{ij})P(p_{ikk'}|\mathbf{z}_{ij},g_{ij1},g_{ij2})}{\sum_{g_{ij1}=0}^{1}\sum_{g_{ij2}=0}^{1}L(g_{ij}=\{g_{ij1},g_{ij2}\}|x_{ij})P(p_{ikk'}|\mathbf{z}_{ij},g_{ij1},g_{ij2})} \tag{7}$$

where $L(g_{ij}|x_{ij})$ gives the pre-calculated likelihood of genotype $g_{ij} \in \{00, 01, 10, 11\}$, with first allele copy g_{ij1} and second allele copy g_{ij2}, given the observed sequence data x_{ij}. $P(p_{ikk'}|\mathbf{z}_{ij}, g_{ij1}, g_{ij2}) = p_{ik}^{g_{ij1}}(1-p_{ik})^{1-g_{ij1}}p_{ik'}^{g_{ij2}}(1-p_{ik'})^{1-g_{ij2}}$ is the product of the allele frequencies of the first and second allele copies of genotype g_{ij} in population $k = z_{ij1}$ and $k' = z_{ij2}$, respectively.

2. Update $\mathbf{z}$ (for all $j \neq f$):

$$P(z_{ijkk'} = 1|g_{ij}, \mathbf{p}_i, \mathbf{Q}_j) = \frac{Q_{jkk'}P(p_{ikk'}|g_{ij})}{\sum_{k=1}^{K}\sum_{k'=1}^{K}Q_{jkk'}P(p_{ikk'}|g_{ij})}, \tag{8}$$

where $P(p_{ikk'}|g_{ij}) = p_{ik}^{g_{ij1}}(1-p_{ik})^{1-g_{ij1}}p_{ik'}^{g_{ij2}}(1-p_{ik'})^{1-g_{ij2}}$ is the product of the allele frequencies of the first and second allele copies of genotype g_{ij} in population k and k', respectively.

3. Update $\mathbf{p}$:

$$P(p_{ik}|\mathbf{z}_i, \mathbf{g}_i, F_k, \pi_i) \sim$$
$$\text{Beta}\left(\pi_i\left(\tfrac{1}{F_k}-1\right)+n_{ijk1}, (1-\pi_i)\left(\tfrac{1}{F_k}-1\right)+n_{ijk0}\right) \tag{9}$$

where

$$n_{ijk1} = \sum_j\sum_a \begin{cases} g_{ija} & \text{when } k = z_{ija}, \\ 0 & \text{when } k \neq z_{ija} \end{cases} \tag{10}$$

and

$$n_{ijk0} = \sum_j\sum_a \begin{cases} 1-g_{ija} & \text{when } k = z_{ija}, \\ 0 & \text{when } k \neq z_{ija} \end{cases} \tag{11}$$

give the counts of the alternative and reference allele copies assigned to have ancestry in population k, for all $j = r$ or m.

4. Update $\boldsymbol{\pi}$:

Propose a new π_i' from

$$\pi_i'|\pi_i \sim \text{Uniform}(\pi_i - 0.1, \pi_i + 0.1), \tag{12}$$

and accept π_i' as new update for π_i with probability $\min(1, r)$ if $0 < \pi_i' < 1$, with

$$r = \frac{P(\alpha|\pi_i')}{P(\alpha|\pi_i)} \prod_k \frac{P(\pi_i', \theta_k|p_{ik})}{P(\pi_i, \theta_k|p_{ik})}, \tag{13}$$

where

$$P(\pi_i, \theta_k|p_{ik}) = \frac{p_{ik}^{\pi_i\theta_k - 1}(1 - p_{ik})^{(1-\pi_i)\theta_k - 1}}{\text{Beta}(\pi_i\theta_k, (1 - \pi_i)\theta_k)} \tag{14}$$

and

$$P(\alpha|\pi_i) = \frac{\pi_i^{\alpha-1}(1 - \pi_i)^{\alpha-1}}{\text{Beta}(\alpha, \alpha)} \tag{15}$$

are calculated from the Beta probability density function, with $\theta_k = \frac{1}{F_k} - 1$ (the probabilities for π_i' are computed in an analogous manner).

5. Update $\boldsymbol{F}$:

Propose a new F_k' from

$$F_k'|F_k \sim \text{Uniform}(F_k - 0.01, F_k + 0.01), \tag{16}$$

and accept F_k' as new update for F_k with probability $\min(1, r)$ if $0 < F_k' < 1$, with

$$r = \prod_i \frac{P(\pi_i, \theta_k'|p_{ik})}{P(\pi_i, \theta_k|p_{ik})}, \tag{17}$$

where $P(\pi_i, \theta_k|p_{ik})$ is given in (14), with $\theta_k = \frac{1}{F_k} - 1$ and $\theta_k' = \frac{1}{F_k'} - 1$.

6. Update α:

Propose a new α' from

$$\alpha'|\alpha \sim \text{Uniform}(\alpha - 20, \alpha + 20), \tag{18}$$

and accept α' as new update for α with probability $\min(1, r)$ if $0 < \alpha' \leq 10000$, with

$$r = \prod_i \frac{P(\alpha'|\pi_i)}{P(\alpha|\pi_i)}, \tag{19}$$

where $P(\alpha|\pi_i)$ is given in (15).

7. Update $\mathbf{Q}$:

The updates for $\mathbf{Q}$ during the MCMC iterations are done differently depending on the family status of individual j.

For mothers, progeny, and the population reference samples, $\mathbf{Q}$ is updated with Gibbs sampling:

$$P(\mathbf{Q}_{j=r,m}|\mathbf{z}_j, \boldsymbol{\gamma}) \sim \text{Dirichlet}(\gamma_{11} + \textstyle\sum_i z_{ij_{11}}, \ldots, \gamma_{KK} + \sum_i z_{ij_{KK}}), \tag{20}$$

$$P(\mathbf{Q}_{j=p}|\mathbf{z}_j, \boldsymbol{\nu}_j, \beta) \sim \text{Dirichlet}(\nu_{j_{11}}\beta + \textstyle\sum_i z_{ij_{11}}, \ldots, \nu_{j_{KK}}\beta + \sum_i z_{ij_{KK}}). \tag{21}$$

We need a Metropolis-Hastings algorithm to update $\mathbf{Q}_{j=f}$. The description of the formulas involves some indexing: For each $j \neq r$, the family membership and status (m, f, or p) is indexed as follows: the mother and father of progeny j, or the female or male mates of a parent j, are indexed as $\overset{(j)}{m}$ and $\overset{(j)}{f}$, and the progeny of a parent j is indexed with $\overset{(j)}{p}$. For each $j = f$ with current $\mathbf{Q}_j$, we propose a new $\mathbf{Q}'_j$ from

$$P(\mathbf{Q}'_j|\mathbf{Q}_j) \sim \text{Dirichlet}(\rho Q_{j_{11}}, \ldots, \rho Q_{j_{KK}}), \tag{22}$$

with $\rho = 200$ being a constant precision scalar. We then calculate the new expected admixture proportion of the progeny of j with $\boldsymbol{\nu}'^{(j)}_p = f(\mathbf{Q}^{(j)}_m, \mathbf{Q}'_j)$ using (4) and (5) and accept $\mathbf{Q}'_j$ as new update for $\mathbf{Q}_j$ with probability $\min(1, r)$, with

$$r = \frac{P(\mathbf{Q}'_j|\boldsymbol{\gamma})P(\mathbf{Q}^{(j)}_p|\boldsymbol{\nu}'^{(j)}_p\beta)P(\mathbf{Q}_j|\rho\mathbf{Q}'_j)}{P(\mathbf{Q}_j|\boldsymbol{\gamma})P(\mathbf{Q}^{(j)}_p|\boldsymbol{\nu}^{(j)}_p\beta)P(\mathbf{Q}'_j|\rho\mathbf{Q}_j)}, \tag{23}$$

with

$$P(\mathbf{Q}_j|\boldsymbol{\gamma}) = \frac{1}{\text{Beta}(\boldsymbol{\gamma})} \prod_{k*=1}^{KK} Q_{jk*}^{\gamma_{k*}-1}, \tag{24}$$

$$P(\mathbf{Q}_p^{(j)}|\boldsymbol{\nu}_p^{(j)}\beta) = \frac{1}{\text{Beta}(\boldsymbol{\nu}_p^{(j)}\beta)} \prod_{k*=1}^{KK} Q_{pk*}^{(j)\,\nu_{pk*}^{(j)}\beta-1}, \tag{25}$$

and

$$P(\mathbf{Q}'_j|\rho\mathbf{Q}_j) = \frac{1}{\text{Beta}(\rho\mathbf{Q}_j)} \prod_{k*=1}^{KK} Q'^{\rho Q_{jk*}-1}_{jk*} \tag{26}$$

calculated from Dirichlet probability density functions ($k*$ indexes the elements of the $K \times K$ matrices of $\mathbf{Q}$, $\boldsymbol{\nu}$, and $\boldsymbol{\gamma}$). Following each updating step of $\mathbf{Q}_j$, the marginal parameter $\mathbf{q}_j$ is calculated for all j as in (4).

8. Update β:

Propose a new β' from

$$\beta'|\beta \sim \text{Uniform}(\beta - 10, \beta + 10), \tag{27}$$

and accept β' as new update for β with probability $\min(1, r)$ if $0 < \beta' < 500$, with

$$r = \prod_j \frac{P(\mathbf{Q}_j|\boldsymbol{\nu}_j\beta')}{P(\mathbf{Q}_j|\boldsymbol{\nu}_j\beta)} \tag{28}$$

for all $j = p$, with $P(\mathbf{Q}_j|\boldsymbol{\nu}_j\beta)$ given in (25).

9. Update $\boldsymbol{\gamma}$ (for all $j \neq p$):

In our current model, all elements $\gamma_{kk'}$ of $\boldsymbol{\gamma}' = (\gamma'_{11}, \ldots, \gamma'_{KK})$ are identical. We therefore propose new $\boldsymbol{\gamma}'$ by proposing one of its elements $\gamma_{kk'}$ (or γ' for brevity) from

$$\gamma'|\gamma \sim \text{Uniform}(\gamma - 0.05, \gamma + 0.05), \tag{29}$$

and accept γ' as new update for γ with probability $\min(1, r)$ if $0 < \gamma' \leq 10$, with

$$r = \prod_j \frac{P(\mathbf{Q}_j|\boldsymbol{\gamma}')}{P(\mathbf{Q}_j|\boldsymbol{\gamma})}, \tag{30}$$

with $P(\mathbf{Q}_j|\boldsymbol{\gamma})$ given in (24).

Estimation of genetic ancestry without family information

As ancestry estimates could potentially be affected by using different priors for adults ($P(\mathbf{Q}_{m,f,r}|\boldsymbol{\gamma})$) and progeny ($P(\mathbf{Q}_p|\boldsymbol{\nu},\beta)$), we additionally ran our model without providing family information. We ran our model by coding all individuals as reference samples, using the same settings as above (5 independent chains combined; 11,976 or 5,226 SNPs). Genetic ancestry for fathers cannot be estimated in that case.

We obtained very similar results for admixture proportions q and inter-source ancestry Q_{12} (Fig. S10; Data S1), indicating that different prior constructions had little effect on our findings. However, results deviated for one of the families, F039 (see Fig. S10). Without family information, ancestries for the maternal tree and some of her progeny were shifted toward pure *P. alba* (results for maternal tree F039, 11,976 SNPs; with family information, $q = 0.942$, $Q_{12} = 0.116$; without family information, $q = 0.986$, $Q_{12} = 0.027$). Although the deviation in ancestry estimates for some of F039's progeny could have resulted from the usage of different priors for $\mathbf{Q}$, the fact that ancestry for maternal tree F039 also differed makes it more likely that F039 has an unusual multilocus genotype, or that the difference arose from an unusual genome-wide ancestry compared to the remainder of the population. In the original model, population allele frequencies were only updated from reference samples and maternal trees, but not progeny, to avoid influence of related individuals on allele frequency estimates. In the model without family information, allele frequencies were estimated from all individuals, and thus data from related individuals could potentially have influenced population allele frequencies and ancestry estimates. Likewise, the prior on ancestry, $\boldsymbol{\gamma}$, was updated from all samples and fathers but excluding progeny in the original model, thus treating progeny as reference samples could also have influenced ancestry estimates.

No matter what the true ancestries are, only 20 out of 483 progeny from a single family

were marginally affected. In addition, the results presented in Table 1 (main article) remain unchanged, as individuals were classified as *P. alba* with $q \geq 0.9$.

We further explored the effect of family data on model performance by coding all individuals as reference samples for simulated data sets. Accuracies in ancestry estimates were very similar with and without providing family data to the model (Fig. S2, S3). As in the original model, in the model without family data ancestry estimates were less accurate for $F = 0.1$. Ancestry estimates also deviated most between the two models for small F (Fig. S11, S12). Providing family data improved model performance for ancestry means (quantified by root-mean-square deviation and correlation), but the true values were less often included in equal tail credible intervals (probably because credible intervals were also narrower).

Simulations of individual data

We used simulated data sets to evaluate how well our model could recover the genetic ancestry of fathers, given the genetic data for reference samples, mothers, and progeny. Overall the simulations were meant to emulate the type of empirical data we have in this study. We simulated 15,000 loci distributed evenly across 20 chromosomes as simplification of the 19 *Populus* chromosomes. Population allele frequencies $\mathbf{p}$ were drawn independently for two parental populations from a Beta distribution with parameters $\text{shape}_1 = \boldsymbol{\pi}(1/F - 1)$ and $\text{shape}_2 = (1 - \boldsymbol{\pi})(1/F - 1)$. The $(1/F - 1)$ term acts as a precision parameter for the distribution, and F is a variance relative to the common ancestral allele frequency $\boldsymbol{\pi}$ and is analogous to Wright's F_{ST} (as in e.g., Gompert *et al.*, 2012; Buerkle & Gompert, 2013). The ancestral allele frequency $\boldsymbol{\pi}$ was drawn from a symmetrical Beta distribution with shape α. We chose different α depending on F to obtain U-shaped population allele frequency distributions similar to those of the empirical data.

Genotypes for two parental species populations (par0 and par1) and F_1 hybrids were generated by sampling directly from population allele frequencies. We simulated subse-

quent hybrid generations F_2, F_3, first generation backcrosses toward par0 (BC_{1_0}) and par1 (BC_{1_1}) by sampling gametes from previous generations, assuming one recombination event per chromosome per meiosis at a random position. We generated family samples by sampling parental gametes of 17 mothers (three each of par0, par1 and F_1, and two each of F_2, F_3, BC_{1_0} and BC_{1_1}), and 30 fathers (six each of par0, par1 and F_1, and three each of F_2, F_3, BC_{1_0} and BC_{1_1}), to generate a total of 510 progeny. Genotypes of mothers, progeny, and 120 population reference samples (each 40 of par0 and par1, 20 F_1, each 10 of F_2 and F_3) were combined.

We generated genotype likelihoods from genotypes as follows. We sampled sequence reads from genotypes with sequence depth drawn from a negative binomial distribution with size 5 and mean of 1, 3 or 8. A negative binomial distribution conformed well to our empirical sequence data and has been used in other studies (e.g. Guenther & Coop, 2013), and allowed zero sequence depth at individual loci (i.e. missing data). The number of reads of alternative alleles were then drawn from a binomial distribution with size given by sequence depth, and probability given by the number of alternative alleles at the simulated genotype divided by two (i.e., 0, 0.5 or 1). Second, we computed genotype likelihoods from a binomial distribution with probabilities 0 for reference allele homozygotes, 1 for alternative allele homozygotes, and 0.5 for heterozygotes, with the number of trials and successes given by sequence depth and reads obtained in the first step. We only kept loci with a minor allele frequency ≥ 0.05, and randomly selected 5,000 loci per data set for subsequent analyses.

We explored nine different simulation settings that are likely to influence the information content of the data sets, namely the combinations of $F = \{0.1, 0.5, 0.8\}$, and mean sequence coverage of $\{1, 3, 8\}$. Details on the settings are provided in Table S3. Simulations were done in `R` (R Development Core Team, 2012), and subsequently analyzed in the same way as the empirical data. `R` simulation code is available at Dryad (doi:10.5061/dryad.kh7sc).

Supplemental References

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Galassi M, Davies J, Theiler J, *et al.* (2009) *GNU Scientific Library: Reference Manual*. Network Theory Ltd.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Guenther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Lindtke D, Buerkle CA, Barbará T, *et al.* (2012) Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in *Populus*. *Molecular Ecology*, **21**, 5042–5058.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

The HDF5 Group (2010) *Hierarchical data format version 5, 2000-2010*. http://www.hdfgroup.org/HDF5.

Supplemental Tables and Figures

Table S1: Statistics on sequencing depth and quality for the full set of 11,976 SNPs. Per-locus and per-individual sequence depth and genotype quality (GQ) were first averaged for each locus over different subsets of individuals (all, adults, progeny), based on data obtained by running `bcftools`. Summary statistics were then computed over loci. GQ gives the $-10log_{10}$ transformation of the probability that the genotype call is wrong.

Samples	Mean depth	Median depth	Min depth	Max depth	Depth sd	Mean GQ
All	8.53	6.71	3.64	248.53	9.47	27.14
Adults	14.06	11.17	5.58	259.35	12.79	38.03
Progeny	6.96	5.44	2.94	246.25	8.71	24.06

Table S2: Statistics on sequencing depth and quality for the reduced set of 5,226 SNPs. Per-locus and per-individual sequence depth and genotype quality (GQ) were first averaged for each locus over different subsets of individuals (all, adults, progeny), based on data obtained by running `bcftools`. Summary statistics were then computed over loci. GQ gives the $-10log_{10}$ transformation of the probability that the genotype call is wrong.

Samples	Mean depth	Median depth	Min depth	Max depth	Depth sd	Mean GQ
All	8.15	6.47	3.68	248.53	10.57	26.42
Adults	13.30	10.81	5.58	259.35	13.58	37.27
Progeny	6.69	5.26	2.96	246.25	9.88	23.35

Table S3: Simulation settings. Sim, name of run; N loci, number of loci kept for analysis; N chr, number of simulated chromosomes. Different values for α, F, and sequence coverage were explored.

Sim	N loci	N chr	α	F	Coverage
sim1	5000	20	0.2	0.1	1
sim2	5000	20	0.5	0.5	1
sim3	5000	20	500	0.8	1
sim4	5000	20	0.2	0.1	3
sim5	5000	20	0.5	0.5	3
sim6	5000	20	500	0.8	3
sim7	5000	20	0.2	0.1	8
sim8	5000	20	0.5	0.5	8
sim9	5000	20	500	0.8	8

Table S4: Genetic correlation between maternal samples. Pairwise correlations were calculated from genotype point estimates (computed from genotype likelihoods of `bcftools` output) for 11,976 SNPs. Numbers in bold indicate pairs of trees that are likely to be ramets of the same clone due to their high genetic similarity.

ID	F008	F009	F010	F011	F020	F021	F022	F026	F030	F031	F032	F033	F036	F039	I_345	I_373	I_396
F008	-																
F009	**0.977**	-															
F010	0.568	0.566	-														
F011	0.565	0.562	0.826	-													
F020	0.521	0.524	0.577	0.576	-												
F021	0.171	0.169	-0.093	-0.100	0.205	-											
F022	0.535	0.538	0.579	0.564	0.549	0.197	-										
F026	0.518	0.519	0.573	0.567	0.549	0.195	0.542	-									
F030	0.208	0.207	-0.058	-0.064	0.231	0.649	0.236	0.223	-								
F031	0.173	0.169	-0.092	-0.099	0.204	**0.985**	0.200	0.197	0.648	-							
F032	0.537	0.535	0.576	0.576	0.544	0.202	0.551	0.543	0.226	0.204	-						
F033	0.537	0.533	0.576	0.564	0.552	0.219	0.547	0.539	0.225	0.219	0.533	-					
F036	0.536	0.534	0.565	0.568	0.55	0.178	0.531	0.525	0.215	0.180	0.532	0.531	-				
F039	0.564	0.563	0.797	0.801	0.554	-0.071	0.567	0.568	-0.038	-0.068	0.566	0.557	0.555	-			
I_345	0.348	0.345	0.211	0.210	0.355	0.375	0.351	0.357	0.395	0.373	0.360	0.368	0.359	0.218	-		
I_373	0.229	0.224	-0.046	-0.056	0.252	0.532	0.243	0.240	0.545	0.532	0.241	0.235	0.226	-0.029	0.399	-	
I_396	0.533	0.532	0.564	0.571	0.545	0.176	0.532	0.528	0.215	0.180	0.533	0.528	**0.978**	0.556	0.356	0.225	-

Table S5: Results of two-sided Kolmogorov-Smirnov tests for the equality of the distributions of admixture proportion (q) or inter-source ancestry (Q_{12}), contrasting them between adults (reference samples and maternal trees) and progeny (all progeny or only progeny from families with pure species mothers with $q \leq 0.1$ or $q \geq 0.9$). Tests were conducted for all samples or hybrid samples only ($0.1 < q < 0.9$ or $Q_{12} > 0.1$), or for hybrid adults vs. hybrid progeny from families with pure species mothers only (this latter test was accomplished to investigate the extreme hypothetical case where hybrid mothers produce negligible seed quantities compared to species mothers). Tests were based on data from 11,976 SNPs; if present, identical values were removed, as the test statistic requires continuous distributions. Numbers in square brackets give number of individuals in each class.

Parameter	Comparison	D	p-value
q	All adults [104] vs. all progeny [452]	0.3525	1.494e-09
	Hybrid adults [50] vs. hybrid progeny [335]	0.5899	1.423e-13
	Hybrid adults [50] vs. hybrid progeny (species mothers) [67]	0.4773	1.661e-06
Q_{12}	All adults [95] vs. all progeny [466]	0.4664	2.442e-15
	Hybrid adults [51] vs. hybrid progeny [349]	0.8949	< 2.2e-16
	Hybrid adults [51] vs. hybrid progeny (species mothers) [73]	0.6594	5.598e-13

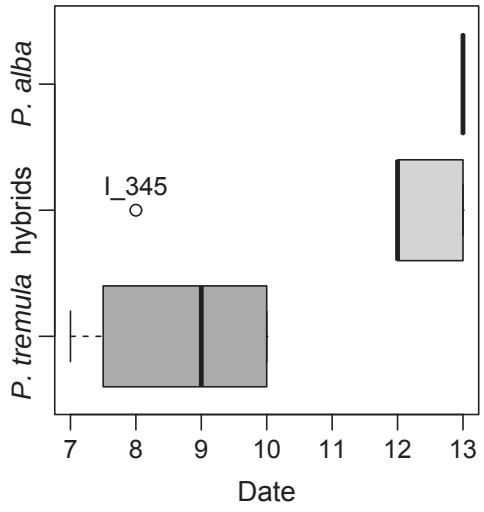

Figure S1: Date of seed collection for open pollinated families, arranged according to the species assignment of the mother. The horizontal axis gives day of April in 2011.

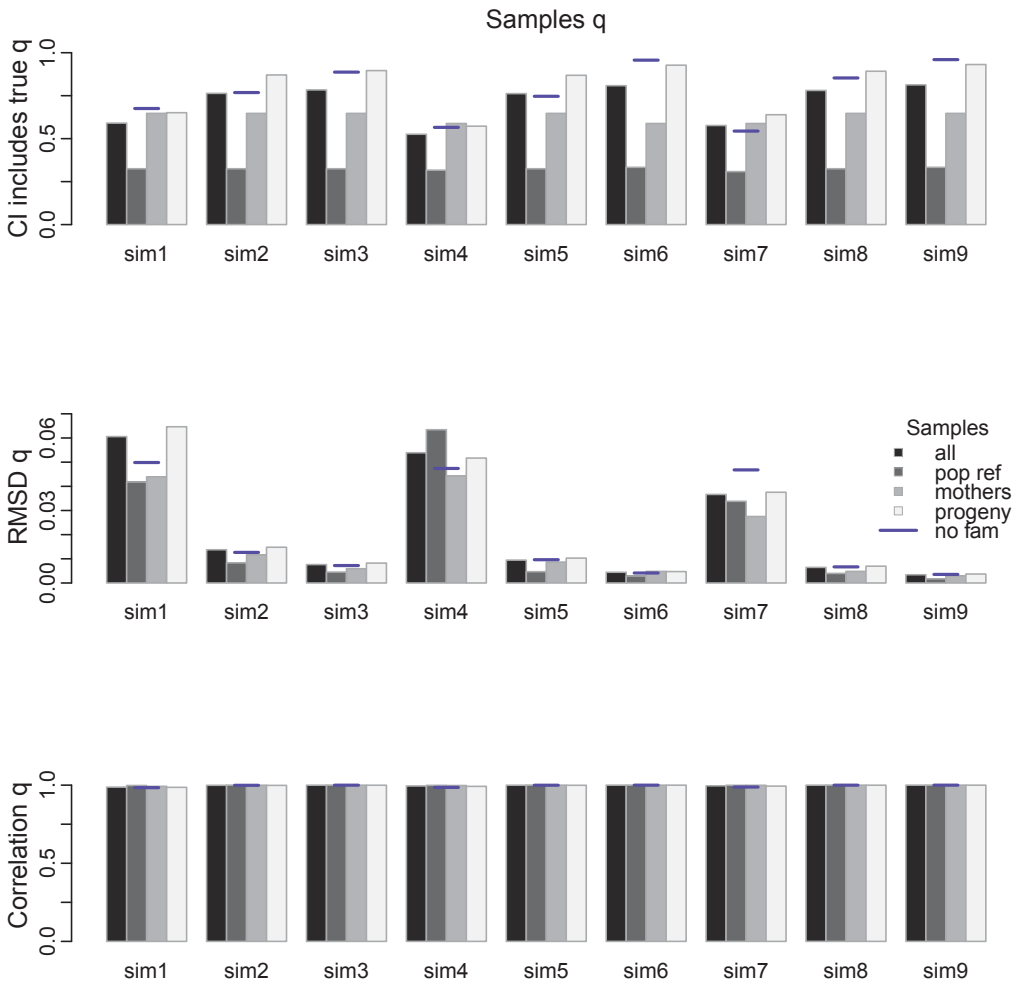

Figure S2: Performance of model for different simulation settings for individual estimates of admixture proportion q for all samples excluding fathers. Top, proportion of individuals where equal tail credible interval (CI) of estimated q includes true q; middle, root-mean-square deviation (RMSD) between true and estimated q; bottom, correlation between true and estimated q. For simulation parameters, see Table S3. Blue horizontal lines indicate model performance when excluding family information.

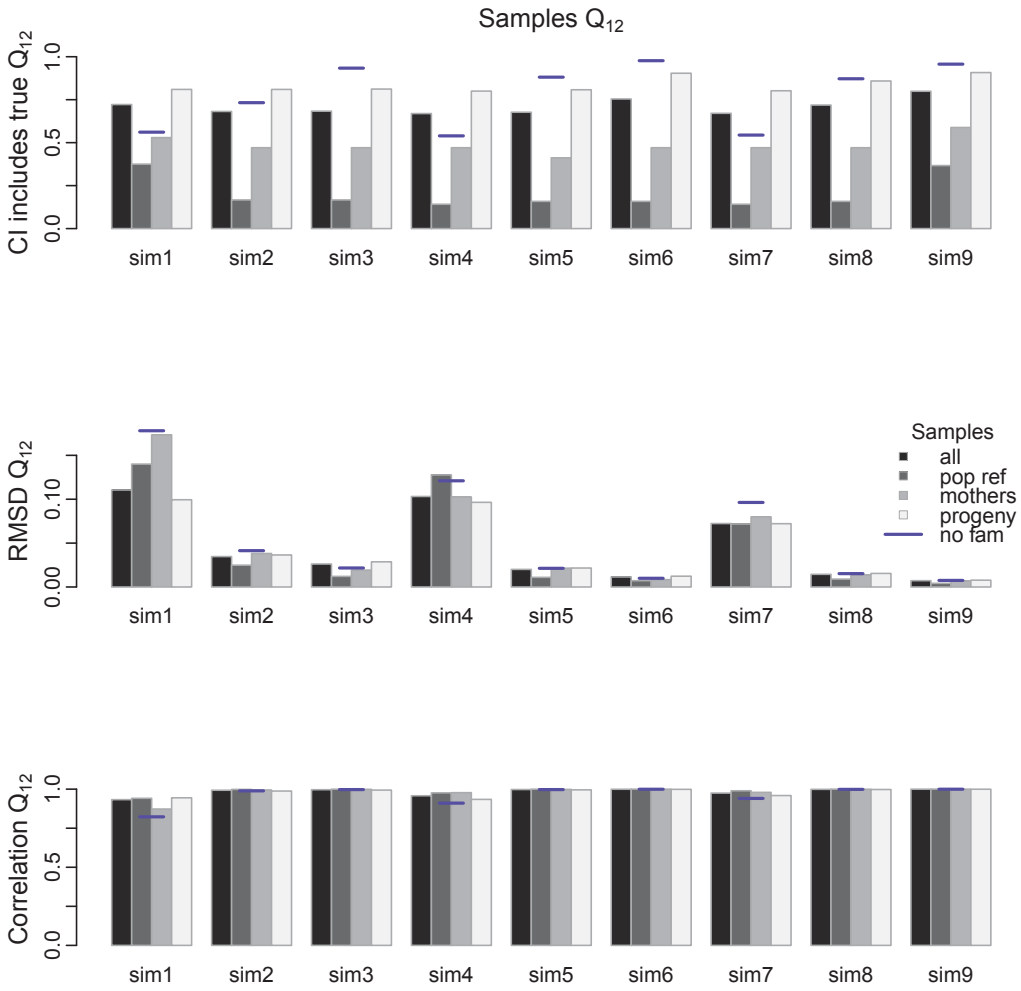

Figure S3: Performance of model for different simulation settings for individual estimates of inter-source ancestry Q_{12} for all samples excluding fathers. Top, proportion of individuals where equal tail credible interval (CI) of estimated Q_{12} includes true Q_{12}; middle, root-mean-square deviation (RMSD) between true and estimated Q_{12}; bottom, correlation between true and estimated Q_{12}. For simulation parameters, see Table S3. Blue horizontal lines indicate model performance when excluding family information.

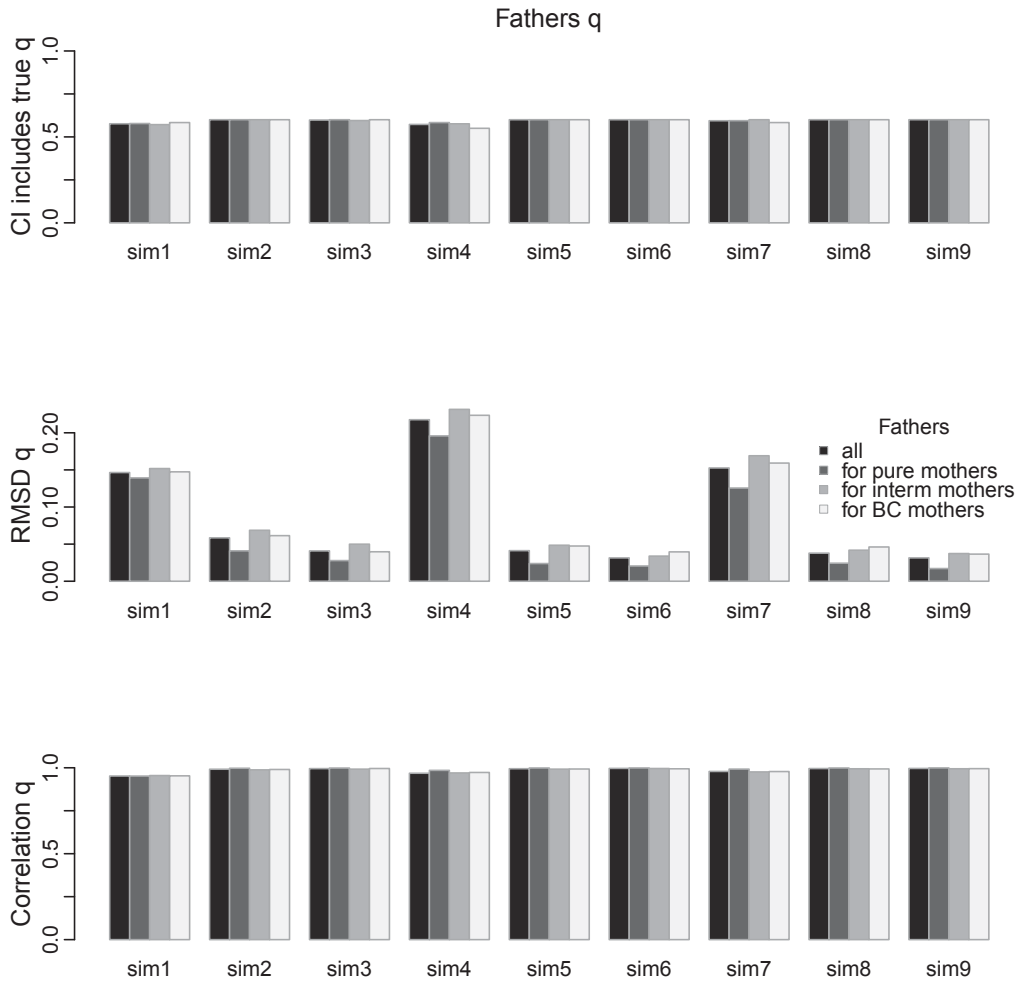

Figure S4: Performance of model for different simulation settings for individual estimates of admixture proportion q of unsampled fathers. Top, proportion of fathers where equal tail credible interval (CI) of estimated q includes true q; middle, root-mean-square deviation (RMSD) between true and estimated q; bottom, correlation between true and estimated q. For simulation parameters, see Table S3.

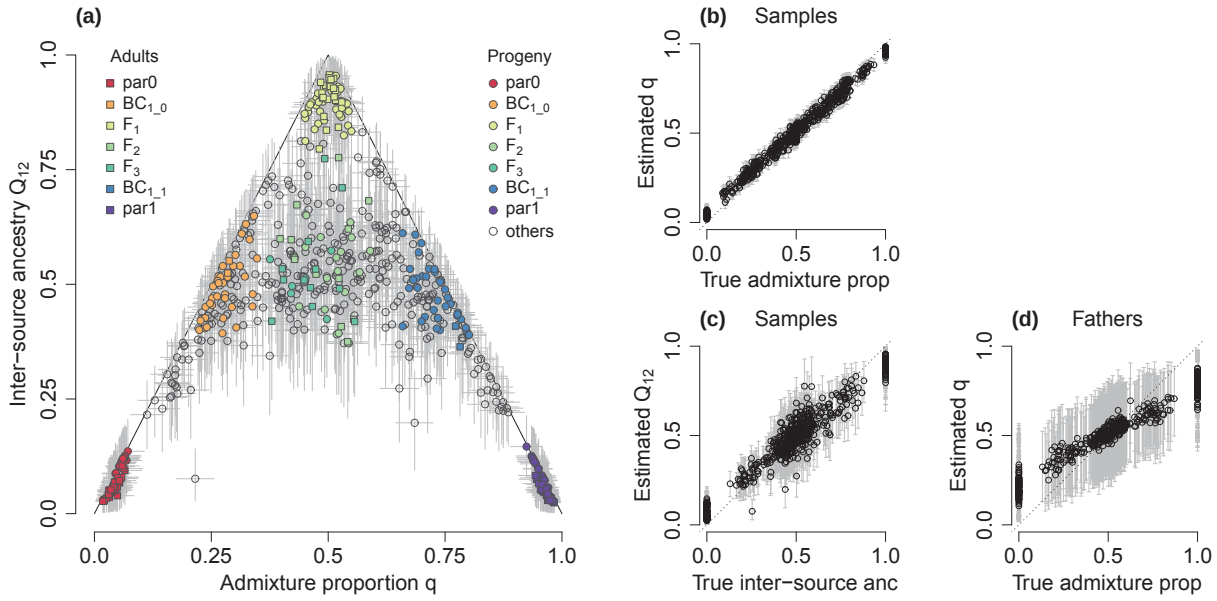

Figure S5: Ancestry estimates for simulated data. The performance of the model and software is shown for sim7 (with $F = 0.1$ and coverage $= 8$). (a) inter-source ancestry (Q_{12}) as a function of admixture proportion (q) for all samples excluding fathers; lines indicate maximum possible Q_{12} given q; (b) comparison of true vs. estimated admixture proportion for all samples excluding fathers; (c) comparison of true vs. estimated inter-source ancestry for all samples excluding fathers; (d) comparison of true vs. estimated admixture proportion for the gametes from fathers. Gray lines show 95% equal tail credible intervals.

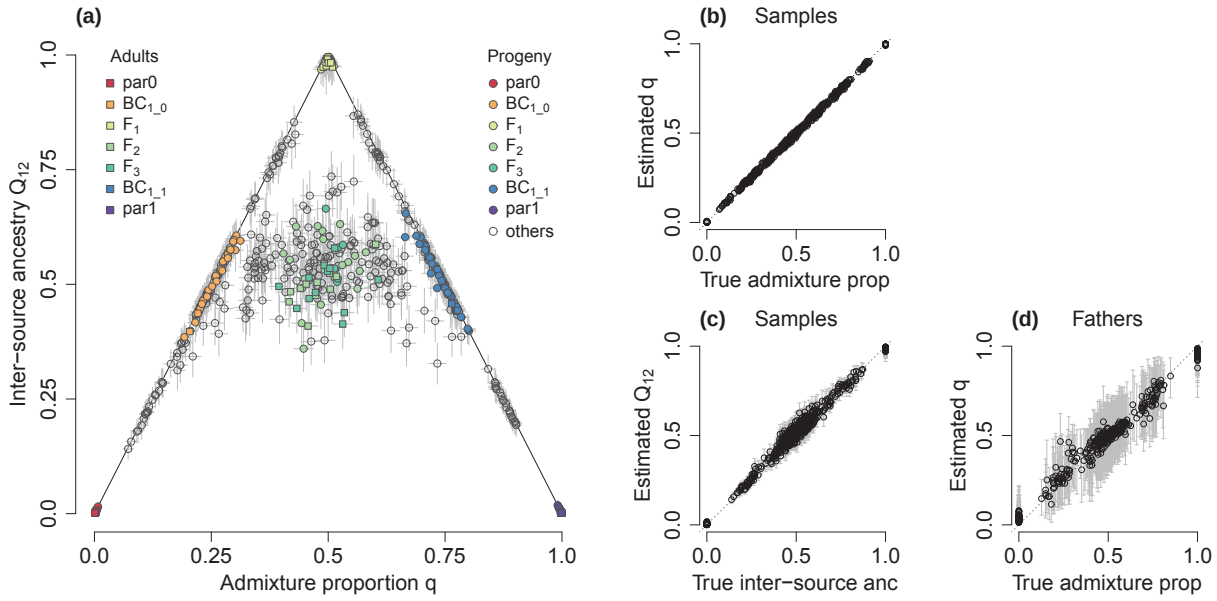

Figure S6: Ancestry estimates for simulated data. The performance of the model and software is shown for sim3 (with $F = 0.8$ and coverage $= 1$). (a) inter-source ancestry (Q_{12}) as a function of admixture proportion (q) for all samples excluding fathers; lines indicate maximum possible Q_{12} given q; (b) comparison of true vs. estimated admixture proportion for all samples excluding fathers; (c) comparison of true vs. estimated inter-source ancestry for all samples excluding fathers; (d) comparison of true vs. estimated admixture proportion for the gametes from fathers. Gray lines show 95% equal tail credible intervals.

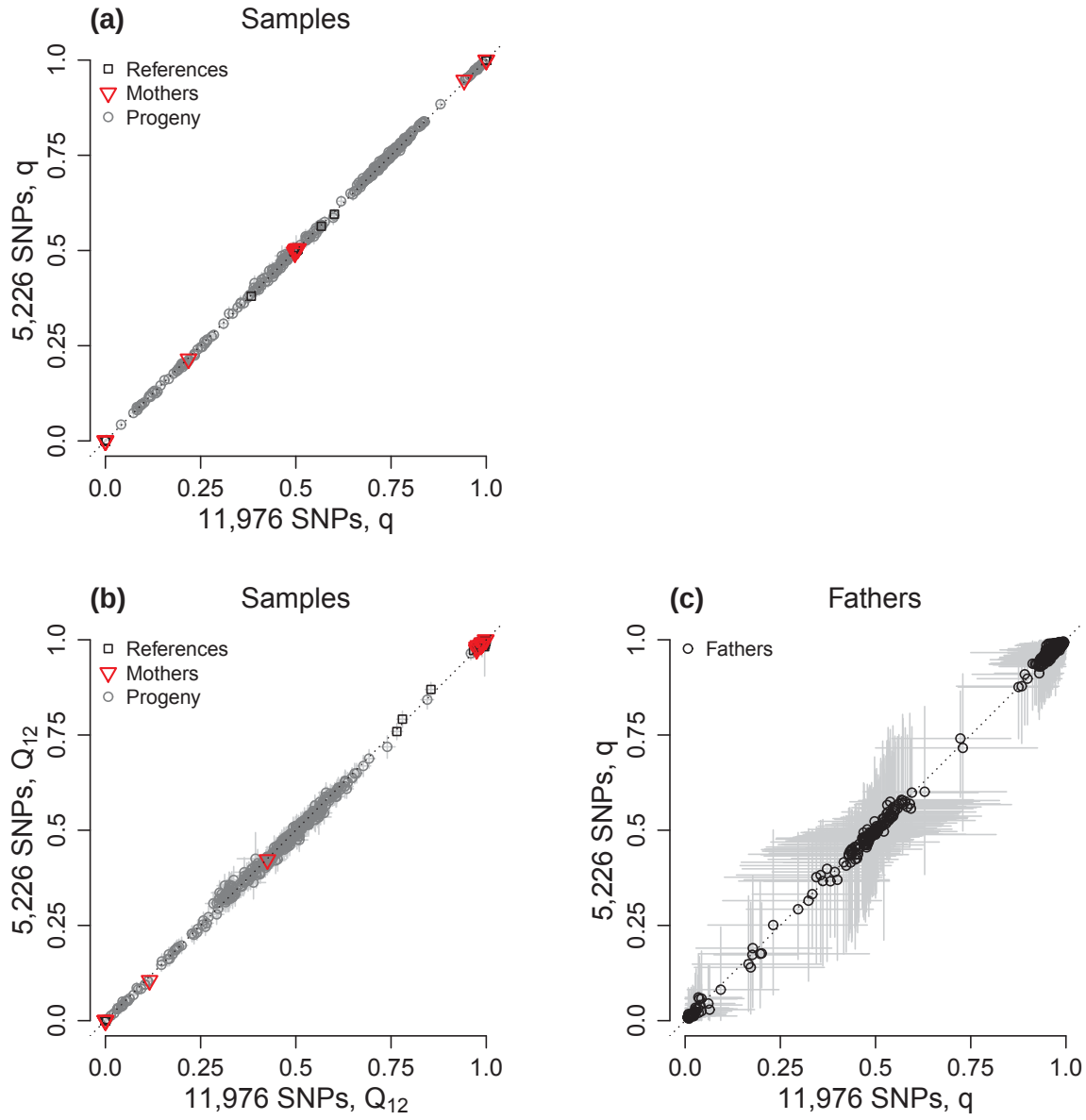

Figure S7: Ancestry estimates obtained for the full (11,976 SNPs) and reduced data set (5,226 SNPs). (a) admixture proportion q for the full vs. reduced data set for all samples excluding fathers; (b) inter-source ancestry Q_{12} for the full vs. reduced data set for all samples excluding fathers; (c) admixture proportion q for the full vs. reduced data set for the gametes from fathers. Gray lines show 95% equal tail credible intervals.

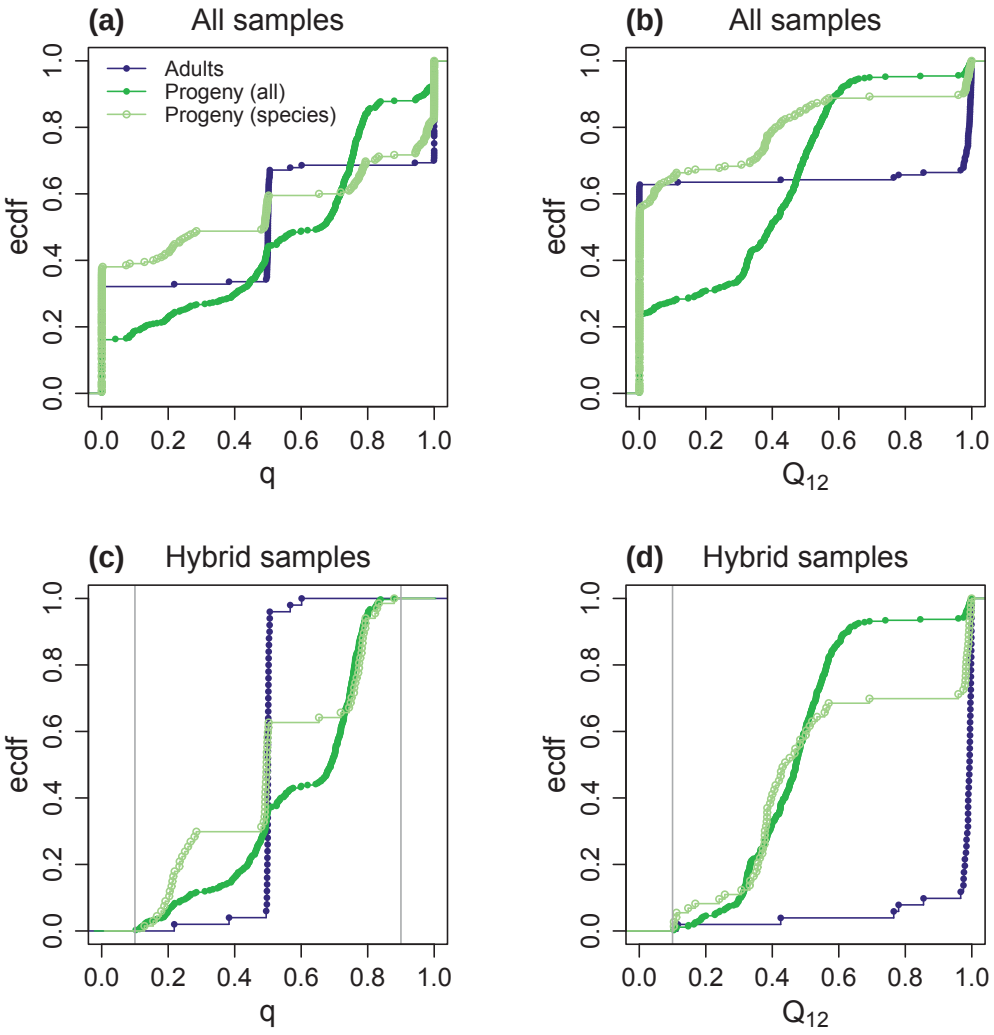

Figure S8: Empirical cumulative distribution functions (ecdf) for admixture proportion (q; left) or inter-source ancestry (Q_{12}; right) for empirical data, based on 11,976 SNPs. The distributions for adults (blue; reference samples and maternal trees) are contrasted to the distributions for all progeny (green, full circles) or progeny from families with pure species mothers only (pale green, open circles; only progeny with mother's $q \leq 0.1$ or $q \geq 0.9$). (a) and (b) show all samples, (c) and (d) only hybrid samples with $0.1 < q < 0.9$ or $Q_{12} > 0.1$; gray vertical lines indicate the thresholds. The distributions for adult and progeny samples were significantly different from each other for all comparisons (Table S5).

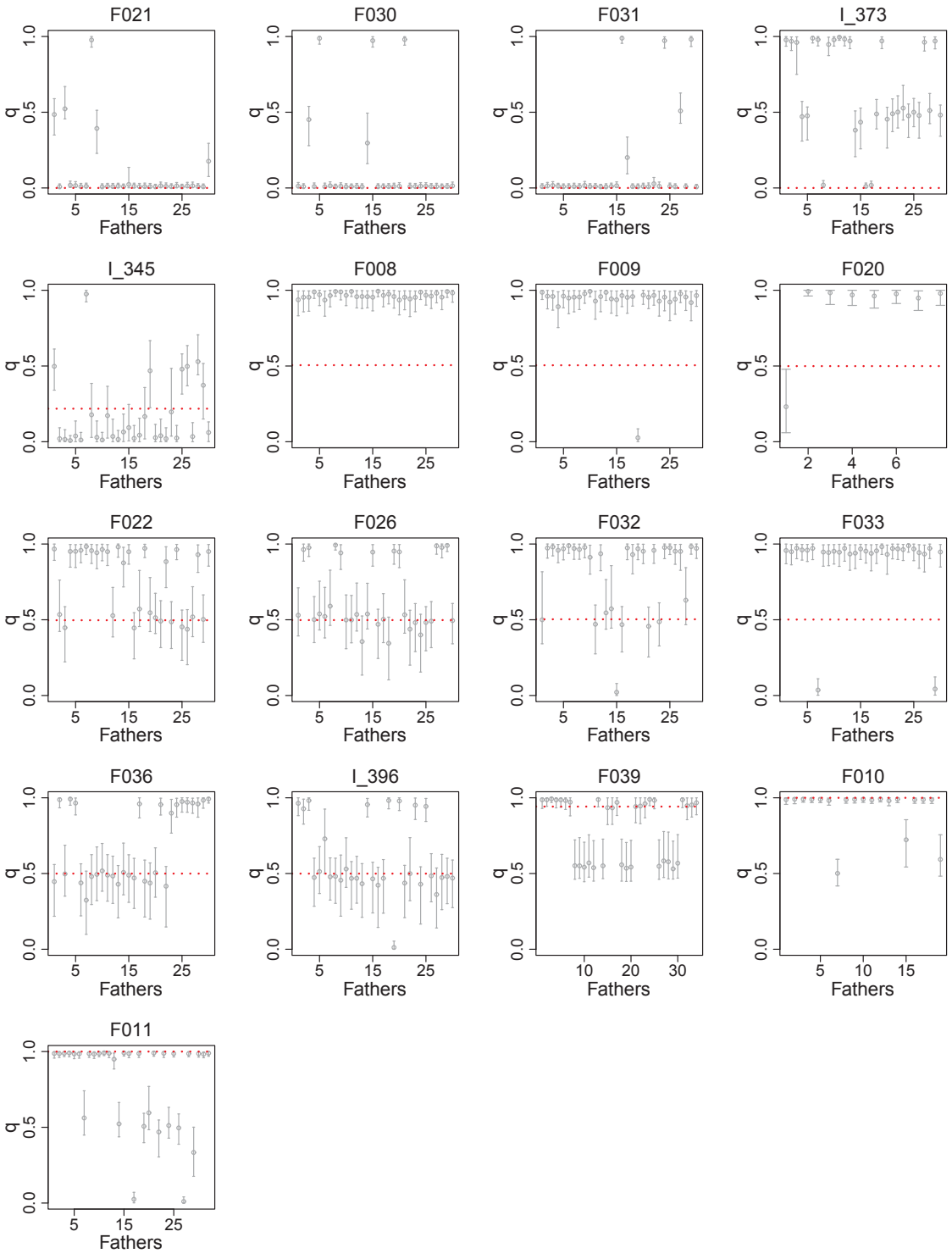

Figure S9: Admixture proportion (q) for the gametes of fathers for each family, based on 11,976 SNPs. Gray lines indicate 95% equal tail credible intervals. Family IDs are given at the top of each plot; the red dotted line shows the admixture proportion of the corresponding maternal parent.

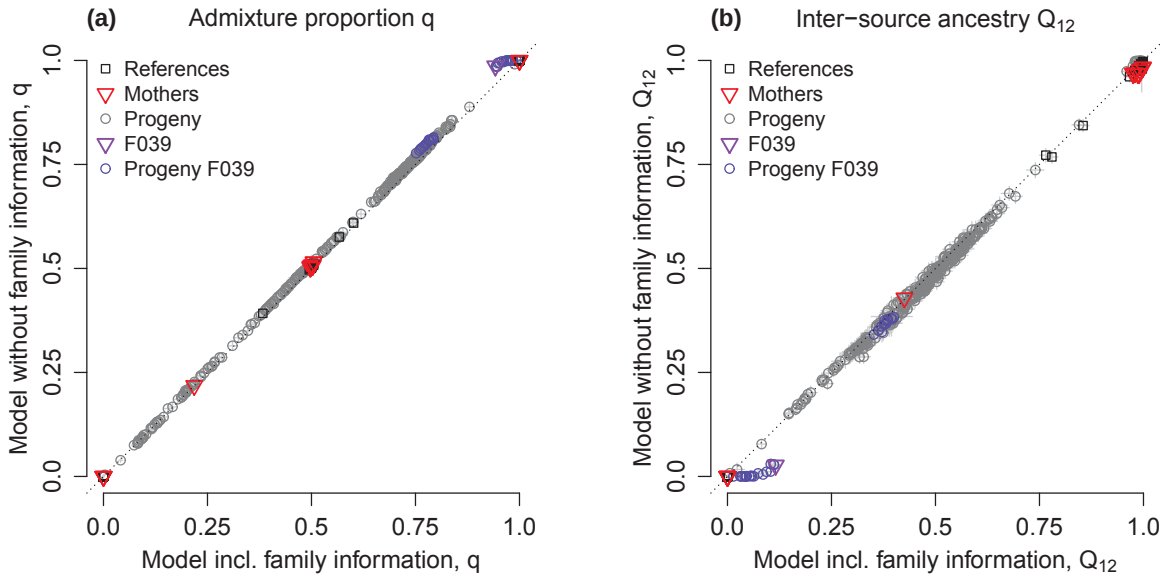

Figure S10: Ancestry estimates obtained with and without using family information, based on 11,976 SNPs. To investigate the potential influence of different priors on $\mathbf{Q}$ used for adults and progeny, genetic ancestry was additionally estimated without providing family information. Results from both models were very similar, but differed marginally for one of the investigated families (see Supporting text for details and discussion; qualitatively identical results were obtained by using the reduced set of 5,226 SNPs). (a) admixture proportion q; (b) inter-source ancestry Q_{12}. Gray lines show 95% equal tail credible intervals.

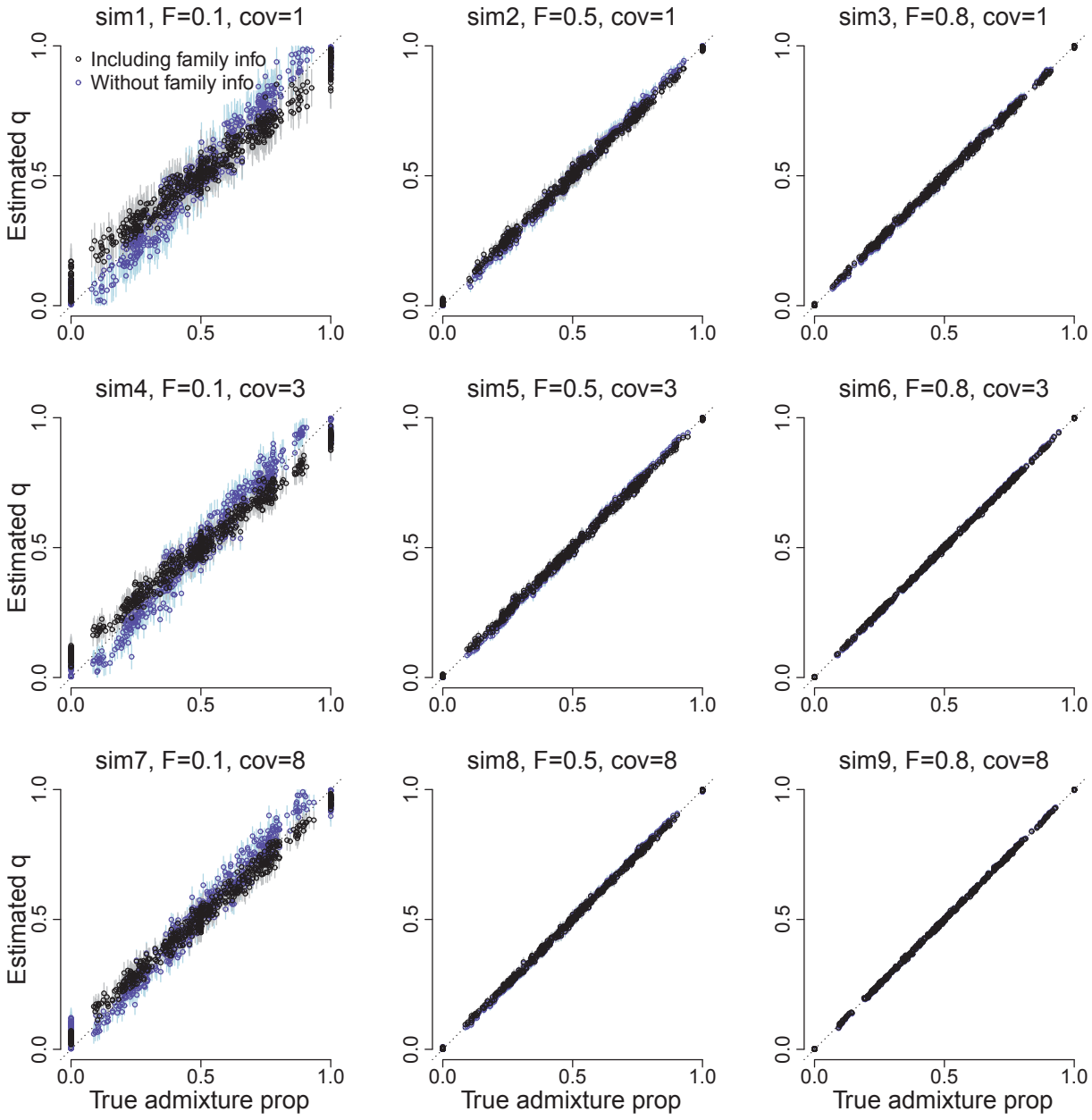

Figure S11: Admixture proportion q obtained with and without using family information, simulated data. Comparison of true vs. estimated admixture proportion for all samples excluding fathers, with providing family information (black circles) or without providing family information (blue circles). Gray and light blue lines show 95% equal tail credible intervals.

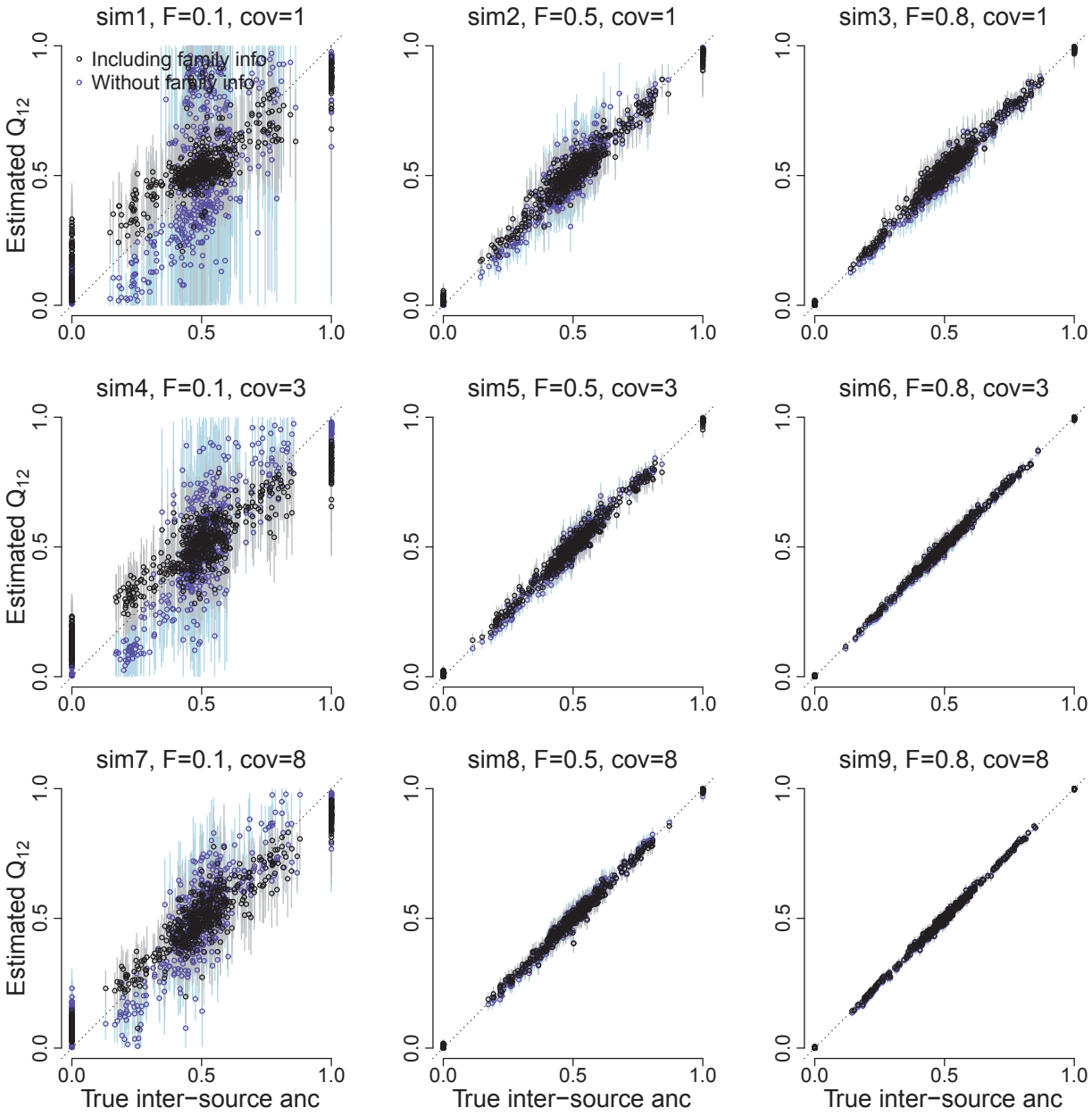

Figure S12: Inter-source ancestry Q_{12} obtained with and without using family information, simulated data. Comparison of true vs. estimated inter-source ancestry for all samples excluding fathers, with providing family information (black circles) or without providing family information (blue circles). Gray and light blue lines show 95% equal tail credible intervals.